DOCUMENT RESUME

ED 290 790 TM 011 030

AUTHOR Brown, R. L.
TITLE Congeneric Modeling of an Interrater Reliability Problem Using Censored Variables.
SPONS AGENCY National Inst. of Mental Health (DHHS), Rockville, MD.
PUB DATE Oct 87
GRANT 1-R01-MH40886-01
NOTE 44p.; Paper presented at the Annual Meeting of the American Evaluation Association (Boston, MA, October 15-17, 1987).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Interrater Reliability; *Monte Carlo Methods; Sample Size
IDENTIFIERS *Asymptotical Distribution Theory; *Joreskogs Congeneric Test Model; *Tobit Model Analysis

ABSTRACT

This paper explores the use of K. G. Joreskog's (1970) congeneric modeling approach to reliability using censored quantitative variables. Two Monte Carlo studies were conducted. The first explored the robustness of Normal Theory Generalized Least-Squares (NTGLS) estimates for a single-factor congeneric model across several sample sizes (N=25,50,100,400), model loading sizes, and different levels of censoring (0%, 25%, 50%, 75%). The second study compared alternate estimation procedures for different levels of variable censoring (0%, 25%, 50%, 75%) based on a single-factor large loading congeneric model. NTGLS, asymptotically distribution free (ADF), and latent TOBIT estimators were compared as to their efficiency in estimating model parameters. Results of the first study indicate that convergence rate is inversely related to sample size and to the size of the model loadings. Censoring of variables produced the expected negative bias in estimates using NTGLS methods with the magnitude of the bias somewhat robust against sample size variation. Results from the second study indicate that the TOBIT estimates were robust with respect to model rejection and amount of parameter bias across various levels of variable censoring. Both NTGLS and ADF methods proved to be unsatisfactory both in model rejections and percent of bias. Thirteen graphs and three data tables conclude the document. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED290790

CONGENERIC MODELING OF AN INTERRATER RELIABILITY PROBLEM USING CENSORED VARIABLES

by

R. L. Brown, Ph.D.
Associate Research Scientist
Schizophrenia Research Unit
352A/425 Henry Mall
University of Wisconsin
Madison, Wisconsin, USA 53706

A paper presented at the Annual Meeting of The American Evaluation Association, Boston, Massachusetts, October 15 - 17, 1987. This research was supported in part by grant 1 R01 MH40886-01 from the National Institute of Mental Health. I wish to express my appreciation to Mary Ann test (University of Wisconsin) and Bengt Muthén (UCLA) for their support.

TM011 030

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Roger Brown

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

Abstract

This paper explores the use of Jöreskog's (1970) congeneric modeling approach to reliability using censored quantitative variables. Two Monte Carlo studies were conducted. The first study explored the robustness of Normal theory Generalized Least-squares estimates for a single factor congeneric model across a variety of sample sizes (N=25, 50, 100, 400), model loading sizes (large λ = .9 and moderate λ = .6), and different levels of censoring (0%, 25%, 50% 75%). A second study compared alternate estimation procedures for different levels of variable censoring (0%, 25%, 50%, 75%) based on a single factor large loading congeneric model with a sample size of N=100. Both normal theory generalized least-squares (NTGLS) (Jöreskog & Goldberger, 1972; Browne, 1974), asymptotically distribution free (ADF) (Browne, 1982), and latent TOBIT estimators (Muthén, 1985; 1987a; in press) were compared as to their efficiency in estimating model parameters.

Results from study one confirmed previous findings that convergence rate is inversely related to sample size, and to some degree size of the model loadings. Censoring of variables produced the expected negative bias in estimates using NTGLS methods with the magnitude of the bias somewhat robust against sample size variation. Results from study two indicated that the TOBIT estimates were robust with respect to model rejection and amount of parameter bias across various levels of variable censoring. Both NTGLS and ADF methods proved to be unsatisfactory both in model rejections and percent of bias.

Introduction

Several procedures have been proposed for the estimation of interrater reliability, many of which have been based on an analysis of variance approach (Winer, 1962). Though these procedures are used frequently they require acceptance of a number of assumptions (Saal, Downey & Lahey, 1980). One major assumption is that all the measures must have the same unit of measurement, sometimes referred to as "tau equivalency" (Lord & Novick, 1968). In response to this assumption, a covariance modeling procedure, based on Jöreskog's (1970) general model for the analysis of covariance structures has been proposed (Werts, Linn & Jöreskog, 1974; Van Der Kamp & Mellenbergh, 1976).

While the covariance modeling procedure provides details about the assumptions surrounding the use of analysis of variance based reliability measures, it is not without some limitation. One major limitation is that the observed variables must be multivariate normally distributed for the appropriate estimation of parameters, if one is using the typical estimation procedures (e.g., maximum likelihood, or normal theory generalized least squares). This procedure would come under question if one attempted to analyze censored variables (variables that have a high concentration of cases at either end of the distribution).

This paper explores the use of Jöreskog's (1970) congeneric modeling approach to reliability using censored quantitative

variables. Two Monte Carlo studies were conducted. The first study explored the robustness of Normal theory Generalized Least-squares (NTGLS) estimates for a single factor congeneric model across various sample sizes (25, 50, 100, 400), model loadings (large λ = .9 and moderate λ = .6), and levels of variable censoring (0%, 25%, 50% 75%). A second study compared alternate estimation procedures for different levels of variable censoring (0%, 25%, 50%, 75%) for the congeneric model, using NTGLS (Jöreskog & Goldberger, 1972; Browne, 1974), asymptotically distribution free methods (ADF) (Browne, 1982), and latent TOBIT estimators (Muthén, 1985; 1987a; in press). These methods were then compared as to their efficiency in estimating model parameters.

Congeneric Modeling Approach to Reliability

The congeneric approach, based on covariance modeling (Jöreskog, 1970) considers a rater as a test instrument, with a data matrix X(m x k), denoting rater scores for a randomly selected subject group (m), where m = the number of observations (subjects) over (k) raters. Assuming the rows of X are independently distributed, each having a multivariate normal distribution, the congeneric model for a four-indicator reliability model may be written as:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{bmatrix} \eta + \begin{bmatrix} \theta_{\varepsilon_1} \\ \theta_{\varepsilon_2} \\ \theta_{\varepsilon_3} \\ \theta_{\varepsilon_4} \end{bmatrix} = \Lambda \eta + \theta_{\varepsilon} \quad (1)$$

where (λ) and (θ_{ε}) are parameters to be estimated for each (k) rater, with (η) representing the true component and (θ_{ε}) the error component. Assuming uncorrelated measurement errors, with variables that have a zero mean and a factor variance of 1.00, the variance-covariance matrix for this model may be defined as:

$$\Sigma = \begin{bmatrix} \lambda^2_1 + \theta_{\varepsilon_1} & & & \\ \lambda_1\lambda_2 & \lambda^2_2 + \theta_{\varepsilon_2} & & \\ \lambda_1\lambda_3 & \lambda_2\lambda_3 & \lambda^2_3 + \theta_{\varepsilon_3} & \\ \lambda_1\lambda_4 & \lambda_2\lambda_4 & \lambda_3\lambda_4 & \lambda^2_4 + \theta_{\varepsilon_4} \end{bmatrix} \quad (2a)$$

which may be rewritten as:

$$\Sigma = \Lambda \Lambda' + \theta_{\varepsilon} \quad (2b)$$

where θ_ε is an (k x k) diagonal matrix of error variances, and Λ' is a vector of loadings on a single common factor. Using this model, one is able to test the initial assumption that all measures have the same underlying true score (τ). For example, if one considers a situation where four independent assessments of patient symptoms are rated using a continuous symptom scale, one would estimate the Λ loading matrix and the error variances θ_ε. For convenience, as previously mentioned, the variance of the true score may be standardized ($\phi_{11} = 1.00$). This approach will provide two overidentifying restrictions in the model (two degrees of freedom), allowing one to test the following position:

$$H_o: \quad S = \hat{\Sigma}$$

If this null hypothesis is not rejected, individual rater reliability may be estimated as:

$$\hat{\rho}_i = \frac{\hat{\lambda}_i^2}{\hat{\lambda}_i^2 + \hat{\theta}_{\varepsilon_i}} \qquad (3)$$

with composite reliability being estimated as:

$$\hat{\rho}_c = \frac{(\Sigma \hat{\lambda}_i)^2}{(\Sigma \hat{\lambda}_i)^2 + \Sigma \hat{\theta}_{\varepsilon_i}} \qquad (4)$$

Subsets of the Congeneric Model

A major assumption in the use of intraclass correlation and generalizability theory procedures (Cronbach, Rajaratnam & Gleser, 1963) is that all measures have the same unit of measurement, sometimes referred to as being "tau-equivalent" (Lord & Novick, 1968). If this assumption is not met, then the averaging of scores would not be meaningful, and the use of ANOVA based procedures would not be appropriate. Jöreskog (1970) provides a procedure based on a subset of the congeneric model to test for tau-equivalence. This assumes that measures are equivalent if the regression weights λ_k are equal in Λ. In the tau-equivalence model, one would test:

$$H_o: \quad \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4$$

with (½k(k + 1)-(k + 1)) degrees of freedom. A total of (k + 1) parameters would then be estimated. If the tau-equivalency hypothesis is rejected, then ANOVA procedures should be rejected for estimating reliability and/or generalizability procedures (werts, Linn & Jöreskog, 1974). If the hypothesis is not rejected

the parameter estimates may be used to estimate individual reliability:

$$\hat{\rho}_i = \frac{\hat{\lambda}^2}{\hat{\lambda}^2 + \hat{\theta}_{\varepsilon_i}} \qquad (5)$$

and composite reliability:

$$\hat{\rho}_c = \frac{(k\ \hat{\lambda}\ )^2}{(k\ \hat{\lambda}\ )^2 + \Sigma\ \hat{\theta}_{\varepsilon_i}} \qquad (6)$$

It should also be noted that the reliabilities estimated from equation (5) may vary contingent upon differing error variances (θ_ε). The stability of error variances is assumed in the ANOVA procedures, but may be tested using another subset of the congeneric model, sometimes referred to as a "parallel model" (Gulliksen, 1968; Jöreskog, 1970). This model implies equality constraints on both the λ_i's and the θ_{ε_i}'s, so one tests the following:

$$H_{o1}: \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4$$
$$H_{o2}: \theta_{\varepsilon_1} = \theta_{\varepsilon_2} = \theta_{\varepsilon_3} = \theta_{\varepsilon_4}$$

with (½k(k + 1) - 2) degrees of freedom. A total of two parameters will be estimated in this specific model. If the parallel hypothesis is rejected, Werts, et al. (1974) assert that the use of ANOVA based procedures would have underestimated the composite reliability. If the parallel hypothesis is not rejected on may estimate the reliabilities as follows:

$$\hat{\rho}_k = \frac{\hat{\lambda}^2}{\hat{\lambda}^2 + \hat{\theta}_\varepsilon} \qquad (7)$$

and composite reliability:

$$\hat{\rho}_c = \frac{(k\ \hat{\lambda})^2}{(k\ \hat{\lambda})^2 + k\ \hat{\theta}_\varepsilon} \qquad (8)$$

Tests of these hypotheses amount to tests of the validity of specific assumptions about the model, which are not provided in the ANOVA based intraclass reliability estimates (Werts, Linn & Jöreskog, 1974).

While the Jöreskog (1970) procedure seems attractive for answering questions regarding measurement unit equivalency and stability of error variances in reliability models, the procedure has limitations of its own.

Congeneric Model Limitations

The test of the aforementioned hypotheses are somewhat contingent upon the type of parameter estimation procedure used in constructing the measurement model. Typically, normal theory maximum likelihood (NTML) or generalized least-squares (NTGLS) estimation procedures have been recommended, simply because these procedures allow the estimation of standard errors and provide a test statistic (likelihood ratio L^2) for assessing the hypotheses (Jöreskog, 1970). The adequacy of using NTML or NTGLS estimation procedures for estimating such a reliability model must of course be based on a number of conditions. First, sample size seems to be a crucial factor. Monte Carlo studies based on normal theory distribution sampling have indicated that the use of NTML and NTGLS estimation procedures on sample sizes of less than 50 result in model convergence problems, while sample sizes of N=25 result in serious improper solutions (Boomsma, 1982; 1983; 1985; Anderson & Gerbing, 1984). Typically, minimum sample size recommendations for unbiased estimates are usually set at $N \geq 100$.

Secondly, distributions with aberrant skewness and kurtosis values result in higher than usual model rejections (Boomsma, 1983) and poorer parameter estimates. It is with this second limitation that this paper will concentrate.

Censored Continuous Variables

One source of aberrant skewness and kurtosis values in sample variables has been the result of analyzing sample truncated normal distribution variables. Hald (1949) has called such variables "censored" variables, since the population from which the sample variable was drawn is considered normally distributed, and the sample being considered incomplete. More formally, a continuous censored variable may be defined as:

$$y = c_l \quad , \quad \text{if } y^* \leq c_l$$
$$y = y^* \quad , \quad \text{if } c_l < y^* < c_u$$
$$y = c_u \quad , \quad \text{if } y^* \geq c_u \quad ,$$

where c_l and c_u represent known lower and upper censoring constants, which may take on values $-\infty$ to $+\infty$. Censored variables may thus be considered variables that have limited variability with a large proportion of cases occurring at one or both end point of the scale. Muthén (1985; 1987a) has shown that increases in censoring have a substantial effect on the attenuation of the population correlation estimates (Figure 1), which has an effect on model estimates.

Insert Figure 1 about here

Censored variables are common place in the measurement of extreme phenomena, such as in the case of measuring psychotic symptomatology. A major question then becomes, how robust is the use of normal theory estimation procedures, specifically NTGLS, for estimating congeneric reliability models with censored variables? While studies have shown that normal theory methods are fairly robust with regards to mild deviations (Boomsma, 1983; Muthén & Kaplan, 1985) questions remain regarding the robustness of NTGLS to moderate and sever deviations.

Monte Carlo Studies

Study Design 1:

Table 1 gives an overview of the first study's design. In this study only NTGLS estimates were used, and were obtained by using the LISCOMP (Muthén, 1987b) computer program on an IBM-4341 mainframe computer. A standardized population covariance matrix (Σ) was created for the four variable single factor congeneric model using equation (2b). Two separate single factor model population covariance matrices were created. The first matrix (based on a highly reliable model with large loadings (LLM)) was established using; Var(η) = 1.00 and Λ_1 = [.9, .9, .9, .9]. The second matrix (based on a less reliable model using moderate loadings (MLM)) was established using; Var(η) = 1.00 and Λ_2 = [.6, .6, .6, .6]. Variances of the errors in measurement

(uniqueness) were defined for both models as $\theta_\varepsilon = 1 - \lambda^2$. Using equation (3), it may be shown that rater reliability for the congeneric model may be reduced to λ^2. Table 2 provides population values and corresponding standard errors. Each Σ matrix was then used to generate 100 sample covariance (S) matrices for each sample size condition (N=25, 50, 100, 400).

Insert Table 1 about here

Insert Figure 2 about here

Insert Table 2 about here

Study Design 2:

Table 3 provides details about the design of the second Monte Carlo study. In this study three different estimation procedures were considered. First, as previously discussed, NTGLS was estimated over the conditions. Second, in an attempt to reduce the possible inflation in L^2 due to non-normality, Browne's (1982) asymptotically distribution free (ADF)

generalized least-squares estimator was considered. Finally, Muthén's (1985; 1987a) latent TOBIT estimator for censored variables was estimated to: (1) provide a better estimate of the correlation, and (2) deal with non-normality. All estimates were obtained on an IBM-4341 mainframe computer using the LISCOMP (Muthén, 1987) program. At this point in the research, only one sample size (N=100) and one single factor congeneric model (λ = .9, with $\theta_\varepsilon = 1 - \lambda^2$, and Var($\eta$) = 1.00) was considered for the comparison analysis (see table 2, and figure 2). Future studies will explore variations in sample size, loading size and various model (multi-factor).

Insert Table 3 about here

Estimation Procedures

Estimations will be represented by the general family of fit functions for analysis of covariance structures, as proposed by Browne (1984). The most familiar estimation procedure currently used is the NTGLS procedure. It's fitting function is:

$$F_{GLS} = (s - \sigma)' W^{-1} (s - \sigma), \qquad (10)$$

where (s) and (σ) refer to (y) based on a Pearsonian correlation/covariance, with a weight matrix $W = \Sigma \otimes \Sigma$ or $S \otimes S$, where $\Sigma = \Lambda \Phi \Lambda' + \Theta_\varepsilon$, and $\sigma = Vec[\Sigma]$ (Jöreskog & Goldberger, 1972).

The second estimator is Browne's (1982) asymptotically distribution free (ADF) generalized least-squares estimator. The fitting function for this estimator is:

$$F_{ADF} = (s - \sigma)' W^{-1} (s - \sigma), \qquad (11)$$

where:

$$W = S_{ijkl} - S_{ij}S_{kl},$$

with s_{ij} and s_{kl} biased sample covariances, and S_{ijkl} is the fourth order multivariate cumulant (Kendall & Stuart, 1977).

The third estimator to be used is Muthén's (1985; 1987a) latent TOBIT estimator. It has the same fitting function as the NTGLS with the exception that instead of (y) referring to a sample based Pearsonian correlation/covariance, (y^*) refers to a latent TOBIT correlation/covariance, based on the underlying assumption of individual univariate normality. The TOBIT fitting function is:

$$F_{TOBIT} = (s^* - \sigma^*) W^{-1} (s^* - \sigma^*), \qquad (12)$$

where s^* and σ^* refer to y^* TOBIT based correlation/covariances, with a weight matrix of $\Sigma^* \otimes \Sigma^*$ or $S^* \otimes S^*$, where $\Sigma^* = \Lambda \Phi \Lambda' + \theta_\varepsilon$, and $\sigma^* = Vec[\Sigma]$ (Muthén, 1985; Muthén & Kaplan, 1985). Muthén's (1985) TOBIT approach is a two-stage estimation of s^*, where the initial stage is the univariate estimation of μ and σ^2 based on a normal distribution for censored variables (Gupta, 1952). The second stage estimates the covariances by maximum likelihood method from the bivariate information by holding μ and σ^2 constant at the univariate estimated levels (Muthén, 1985). This approach may be visually conceptualized in figure 3, where a latent correlation is estimated for two censored variables.

Insert Figure 3 about here

Results

The results of Monte Carlo study number 1 support previous findings (Anderson & Gerbing, 1984; Boomsma, 1985) that nonconvergence rate is inversely related to sample size at 0% censoring. Figure 4a and 4b show the frequency of nonconvergence

for the NTGLS estimates for the single factor model over two levels model reliability (loading sizes), four levels of sample size and four levels of variable censoring.

Insert Figures 4a and 4b about here

The asymmetry provided in the higher censoring conditions seems to enhance nonconvergence, but more in the larger loading models than the smaller, with the exception of the N=25 condition. Figures 5a and 5b indicate that variable censoring has a dramatic effect on the proportion of congeneric model rejections, though one must be cautioned about the interpretation of the proportion of model rejections for small samples size (e.g., N=25 and N=50).

Insert Figures 5a and 5b about here

For convenience, the amount of bias will be studied in only two of the four λ estimates (λ_1 a noncensored continuous variable, and λ_4 a censored continuous variable). Figures 6a and 6b present the percent of bias in these two estimates over the studied conditions. In the censored variable (λ_4), the percent of

bias seems to be a monotonic function of the amount of variable censor, regardless of the sample size, with slightly stronger bias occurring in the highly reliable model (large loading size). If one disregards the L^2 test on the basis of smaller sample sizes and calculates individual rater reliability, since individual reliability may be reduced to λ^2_i in this example, we may interpret Figure 6a as the amount of bias in the estimate of reliability for rater four.

The amount of bias in the noncensored continuous variable (λ_1) may be considered trivial with a slight increase in model censoring. The bias in the higher reliability model ($\lambda_i = .9$) tended to be negative, whereas the lower reliability model ($\lambda_i = .6$) tended to fluctuate more (Figure 6b).

Insert Figures 6a and 6b about here

A much more dramatic occurrence of negative bias may be seen in figure 7a, the percent of bias in the estimate of the error variance (θ_ε) in the censored variable. Here one may see a strong negative bias even in the low censoring conditions. Sample size, in this example, seems to have little effect on bias once censoring has occurred. Model loading size produced a slight difference in bias, with the less reliable model providing stronger bias. Figure 7b show the percent of bias in the estimate of error variance in the non-censored variable. As seen in figure

7a, the less reliable model (smaller loadings) seems more susceptible to bias in the estimate of error variance.

Insert Figures 7a and 7b about here

Using the results from study number one, a second Monte Carlo study was designed using a single sample size (N=100) on the single congeneric model (see Figure 2). Figure 8 provides the proportion of model rejections for the three estimation procedures (NTGLS, ADF and TOBIT) across the four levels of variable censoring.

Insert Figure 8 about here

The ADF procedure was initially incorporated to deal with the non-normality of the censored models, and the TOBIT procedure was incorporated to deal with both non-normality and the attenuation in the correlation estimates. All three procedures similarly reject few congeneric models at zero censoring, but as one increases censoring the proportion of rejection increases in both NTGLS and ADF. NTGLS procedures in this example are not at all robust at the 50% censoring level, with the ADF procedure

also providing poor results. This may also be seen in figures 9a and 9b.

Insert Figures 9a and 9b about here

Figure 9a shows the percent of bias occurring in the λ_4 loading estimate (one of the censored conditions). Again the percent of bias is monotonically related to the amount of censoring in both NTGLS and ADF estimation procedures. The TOBIT estimates on the other hand, produced very little bias. The percent of bias occurring for λ_1 (noncensored condition) was negligible for the TOBIT and NTGLS estimates, but resulted in larger negative bias with the use of the ADF procedure (Figure 9b).

The percent of bias in the estimate of the error variance also resulted in a strong positive monotonic relationship with the level of censoring for both NTGLS and ADF procedures. TOBIT estimates were consistently positively biased, but at a relatively low level (figure 10a). Figure 10b shows the percent of bias in the estimate of the error variance for the non-censored condition. The most striking result in this figure is the consistent positive bias produced by the TOBIT estimates.

Insert Figures 10a and 10b about here

Discussion

Monte Carlo study number one demonstrated very well the lack of robustness in the NTGLS estimation procedure for retrieving correct λ parameter estimates, estimated error variances θ_ε's, and model fit indices (likelihood ratios (L^2)) at 25% or more censoring. One must also be impressed with the general lack of model convergence with a small sample size (N=25) for such a well defined model. This point in itself may prove to be the limiting factor against the use of congeneric modeling techniques in interrater reliability studies. While not reported in this paper, concern for improper solutions (e.g., negative θ_ε estimates) have also been shown to be related to small sample size, which would cause difficulty in estimating reliability.

One may conclude from the first Monte Carlo study, that the use of Jöreskog's (1970) congeneric modeling approach to interrater reliability estimation using NTGLS methods must be conducted very cautiously. While superior in many ways to the ANOVA approaches, the congeneric model is very susceptible to even mild deviations from normality (e.g., censored variables) and requires a moderately large sampling (at least N=100) for appropriate estimation. If one is not willing to accept these limitations, then one will certainly have to accept limitations in the reliability estimates.

Monte Carlo study number two demonstrated that dealing with non-normality alone (e.g., ADF procedure) is not enough for providing accurate reliability estimates in censored variable models. One must also be concerned with the attenuation in the correlations. This study demonstrated the superiority of the TOBIT estimates in dealing with these two conditions. If one can assume an underlying normality position, then latent correlation based estimates of reliability may prove to be the savior of the congeneric modeling approach to interrater reliability, though sample size is still a limiting factor. While the latent correlation based estimates seem attractive, further Monte Carlo studies need to be conducted before they can be fully accepted.

References

Anderson, J. C., and Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. Psychometrika, 49, 155-173.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In. K. G. Jöreskog & H. Wold (Eds.), Systems under indirect observation: Causality, structure, prediction (Part 1, pp. 149-173) Amsterdam: North-Holland.

Boomsma, A. (1983). On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality. Unpublished doctoral dissertation, University of Groningen, Groningen.

Boomsma, A. (1985). Nonconvergence, improper soluations, and starting values in lisrel maximum likelihood estimation. Psychometrika, 50, 229-242.

Browne, M. W., (1974). Generalized least squares estimates in the analysis of covariance structures. South African Statistical Journal, 8, 1-24. Reprinted in: D. J. Aigner and A. S. Goldberger, (eds.), 1977, Latent variables in socio-economic models. Amsterdam: North-Holland.

Browne, M. W., (1982). Covariance structures. In D. M. Hawkins, (Ed.), Topics in applied multivariate analysis. Cambridge: Cambridge University Press.

Cronbach, L. J., Rajaratnam, N., and Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137-163.

Gupta, A. K. (1952). Estimation of the mean and standard deviation of a normal population from a censored sample. Biometrika, 39, 260-273.

Hald, A. (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. Skand. AktuarTidskr, p. 119.

Jöreskog, K. G., and Goldberger, A. S. (1972). Factor analysis by generalized least squares. Psychometrika, 37, 243-260.

Jöreskog, K. G. (1970). A general method for analysis of covariance structures. Biometrika, 57, 239-251.

Kendall, M., and Stuart, A. (1977). The advanced theory of statistics Vol. 1. (4th ed.). New York: MacMillan.

Lord, F. M., and Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Muthén, B. (1985, July). Tobit factor analysis. Paper presented at the Fourth European Meeting of the Psychometric Society, Cambridge, England.

Muthén, B., and Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. British Journal of Mathematical and Statistical Psychology, 38, 171-189.

Muthén, B. (1987a, April). Advances in factor analysis and structural equation modeling with categorical and other non-normal data. Paper presented at the meeting of Advances in Factor Analysis, University of Wisconsin, Madison, WI.

Muthén, B. (1987b). LISCOMP: Analysis of linear structural relations using a comprehensive measurement model. [Computer program]. Scientific Software, Inc.: Mooresville, IN.

Muthén, B. (in press). Tobit factor analysis. British Journal of Mathematical and Statistical Psychology.

Saal, F. E., Downey, R. G., and Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88, 413-428.

Van Der Kamp, L. J. T., and Mellenbergh, G. J. (1976). Agreement between raters. Educational and Psychological Measurement, 36, 311-317.

Werts, C. E., Linn, R. L., and Jöreskog, K. G. (1974). Intraclass reliability estimates: Testing structural assumptions. Educational and Psychological Measurement, 34, 25-33.

Winer, B. J. (1962). Statistical principles in experimental design. New York: McGraw-Hill.

Table 1

Monte Carlo Study Design Number 1

Estimation procedure	Number of variables	Loading sizes	Sample sizes	Percent[1] of censor	Number of replications
NTGLS	4	$\lambda = .9$	25	0%	100
NTGLS	4	$\lambda = .9$	50	0%	100
NTGLS	4	$\lambda = .9$	100	0%	100
NTGLS	4	$\lambda = .9$	400	0%	100
NTGLS	4	$\lambda = .6$	25	0%	100
NTGLS	4	$\lambda = .6$	50	0%	100
NTGLS	4	$\lambda = .6$	100	0%	100
NTGLS	4	$\lambda = .6$	400	0%	100
NTGLS	4	$\lambda = .9$	25	25%	100
NTGLS	4	$\lambda = .9$	50	25%	100
NTGLS	4	$\lambda = .9$	100	25%	100
NTGLS	4	$\lambda = .9$	400	25%	100
NTGLS	4	$\lambda = .6$	25	25%	100
NTGLS	4	$\lambda = .6$	50	25%	100
NTGLS	4	$\lambda = .6$	100	25%	100
NTGLS	4	$\lambda = .6$	400	25%	100
NTGLS	4	$\lambda = .9$	25	50%	100
NTGLS	4	$\lambda = .9$	50	50%	100
NTGLS	4	$\lambda = .9$	100	50%	100
NTGLS	4	$\lambda = .9$	400	50%	100
NTGLS	4	$\lambda = .6$	25	50%	100
NTGLS	4	$\lambda = .6$	50	50%	100
NTGLS	4	$\lambda = .6$	100	50%	100
NTGLS	4	$\lambda = .6$	400	50%	100
NTGLS	4	$\lambda = .9$	25	75%	100
NTGLS	4	$\lambda = .9$	50	75%	100
NTGLS	4	$\lambda = .9$	100	75%	100
NTGLS	4	$\lambda = .9$	400	75%	100
NTGLS	4	$\lambda = .6$	25	75%	100
NTGLS	4	$\lambda = .6$	50	75%	100
NTGLS	4	$\lambda = .6$	100	75%	100
NTGLS	4	$\lambda = .6$	400	75%	100

[1] Two of the four λ's remained continuous (0% censor) throughout the study for internal comparison.

Table 2

Monte Carlo Study Design Number 2

Estimation procedure	Number of variables	Loading sizes	Sample sizes	Percent[1] of censor	Number of replications
NTGLS	4	$\lambda = .9$	100	0%	100
ADF	4	$\lambda = .9$	100	0%	100
TOBIT	4	$\lambda = .9$	100	0%	100
NTGLS	4	$\lambda = .9$	100	25%	100
ADF	4	$\lambda = .9$	100	25%	100
TOBIT	4	$\lambda = .9$	100	25%	100
NTGLS	4	$\lambda = .9$	100	50%	100
ADF	4	$\lambda = .9$	100	50%	100
TOBIT	4	$\lambda = .9$	100	50%	100
NTGLS	4	$\lambda = .9$	100	75%	100
ADF	4	$\lambda = .9$	100	75%	100
TOBIT	4	$\lambda = .9$	100	75%	100

[1] Two of the four λ's remained continuous (0% censor) throughout the study for internal comparison.

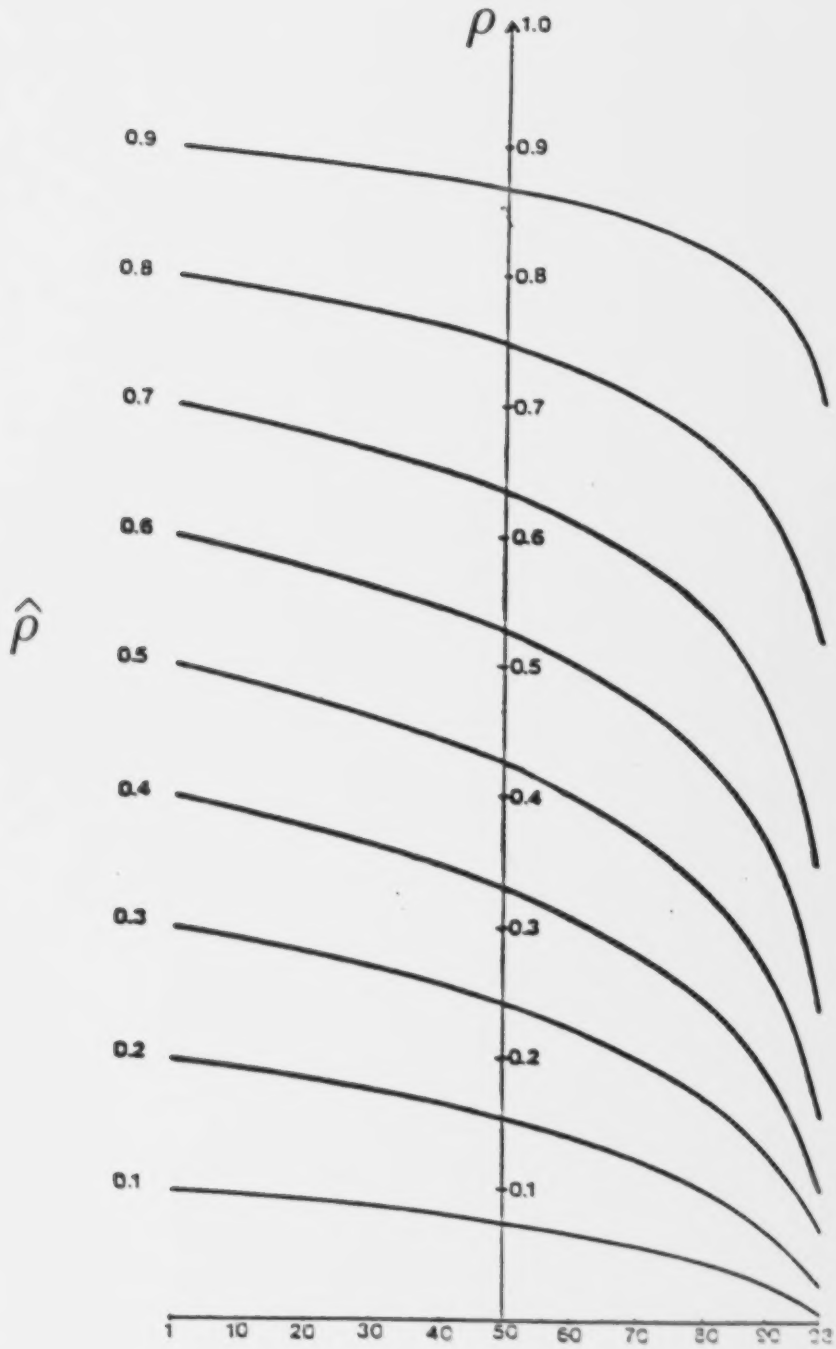
Table 3

Model Population Values and Corresponding Standard Errors

Parameter	Value	Standard Errors			
		N=25	N=50	N=100	N=400
Large Loading					
$\lambda_1\ \lambda_2\ \lambda_3\ \lambda_4$	.900	.160	.112	.079	.039
$\theta_{\varepsilon_1}\ \theta_{\varepsilon_2}\ \theta_{\varepsilon_3}\ \theta_{\varepsilon_4}$	.190	.074	.052	.037	.018
Moderate Loading					
$\lambda_1\ \lambda_2\ \lambda_3\ \lambda_4$	.600	.232	.163	.114	.057
$\theta_{\varepsilon_1}\ \theta_{\varepsilon_2}\ \theta_{\varepsilon_3}\ \theta_{\varepsilon_4}$	.640	.250	.175	.123	.061

Figure 1

Correlation attentuation due to variable censoring[1]



Source: Muthén (1987a).

Figure 2

Single Factor Congeneric Model

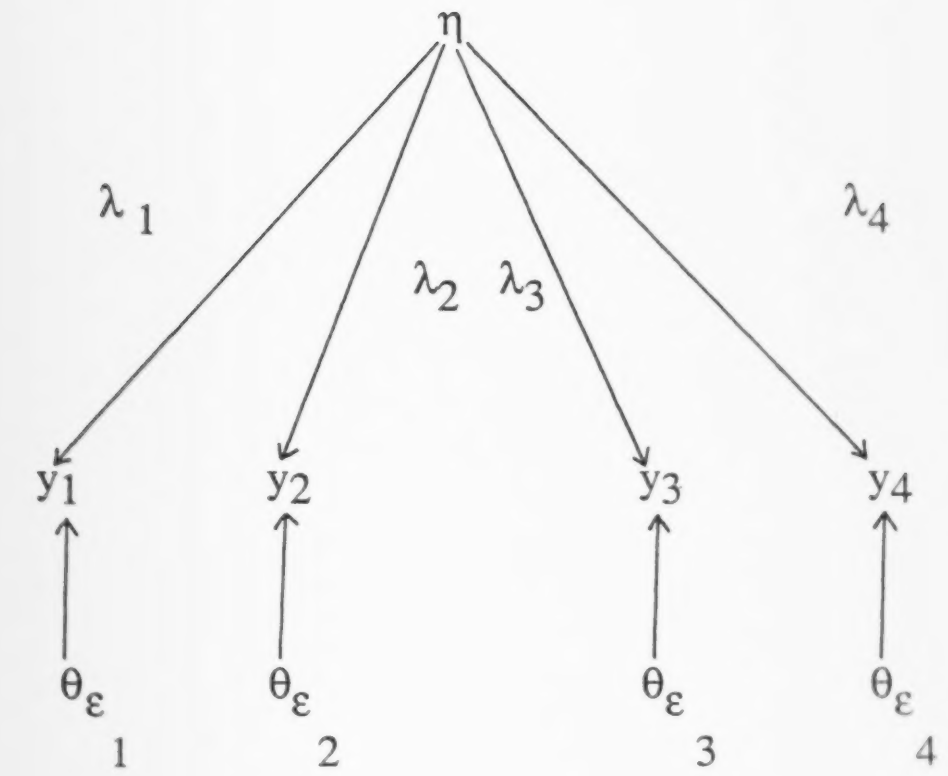


Figure 3

Concept of a latent correlation

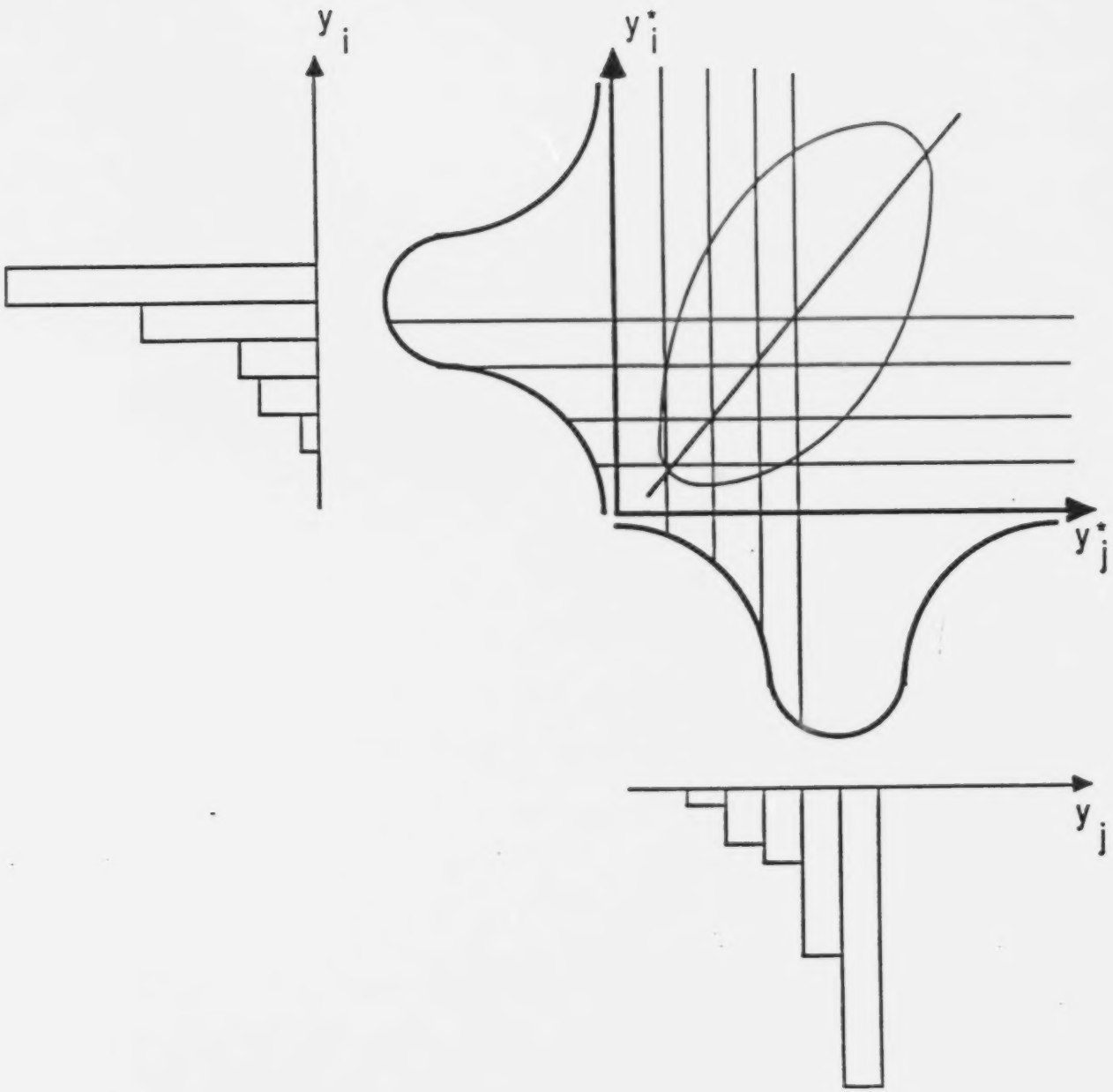
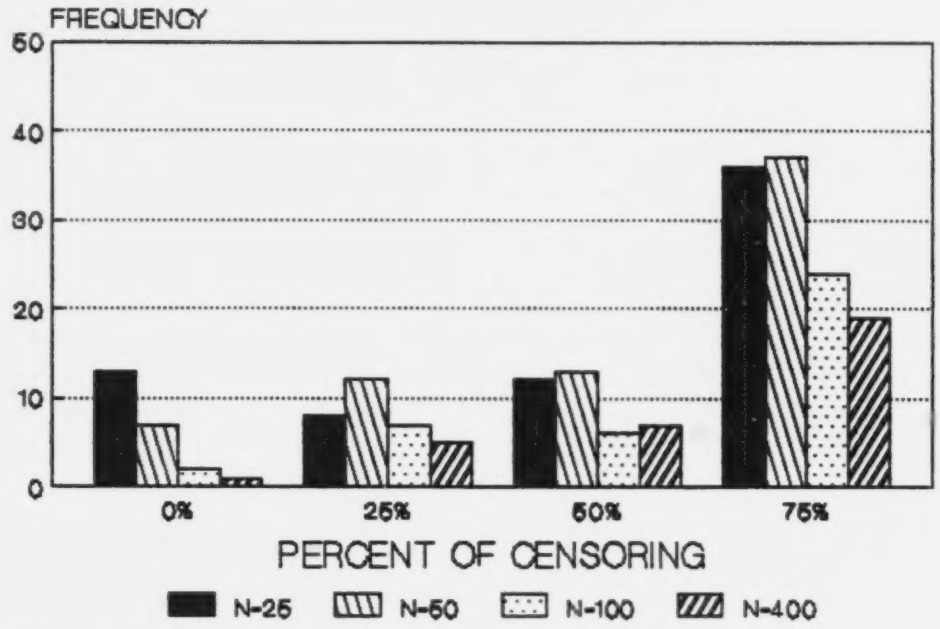
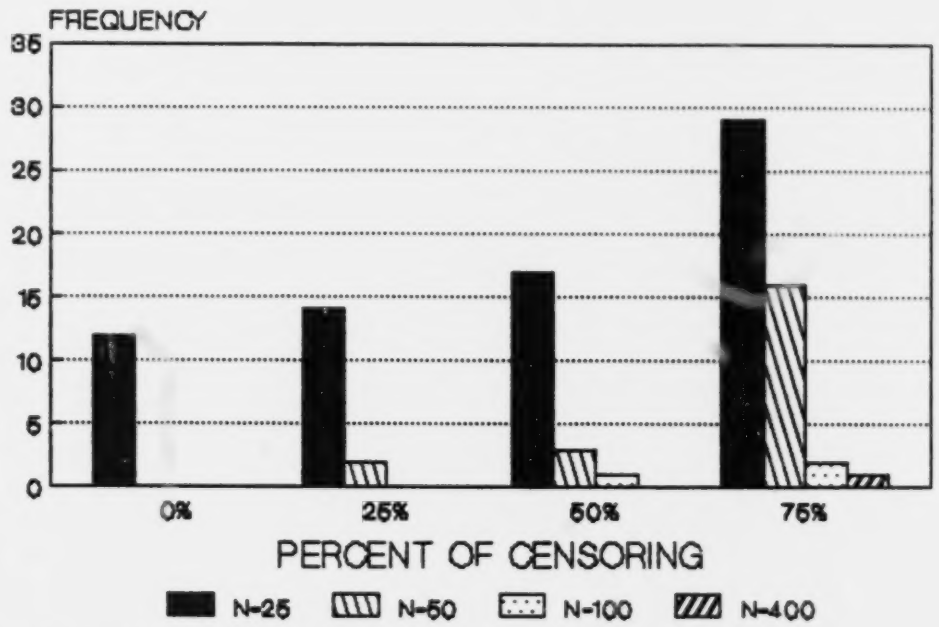


FIGURE 4a
FREQUENCY OF NONCONVERGENCE



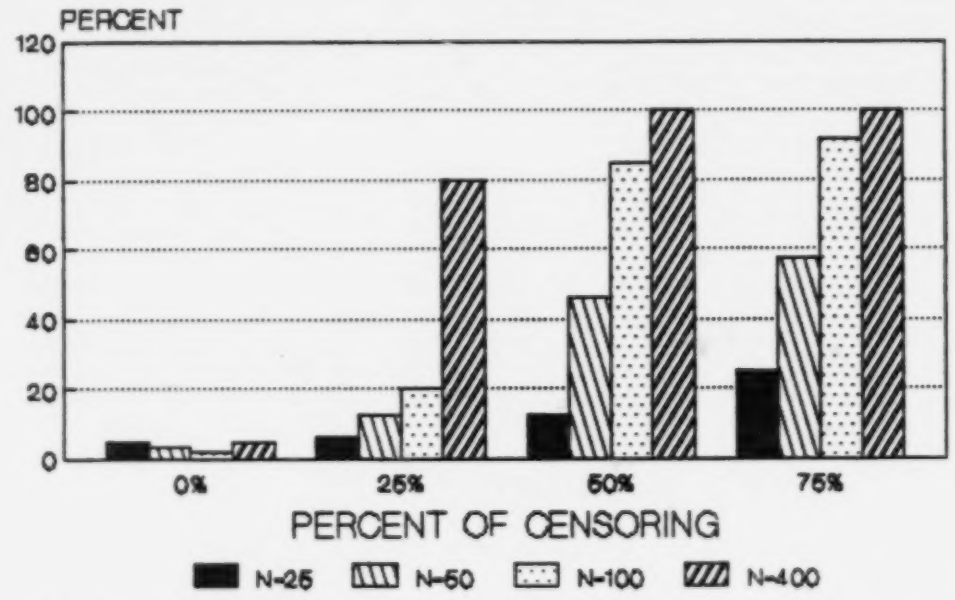
HIGH RELIABILITY MODEL Lambda = .9
NTGLS ESTIMATION

FIGURE 4b
FREQUENCY OF NONCONVERGENCE



MODERATE RELIABILITY MODEL Lambda= .6
NTGLS ESTIMATION

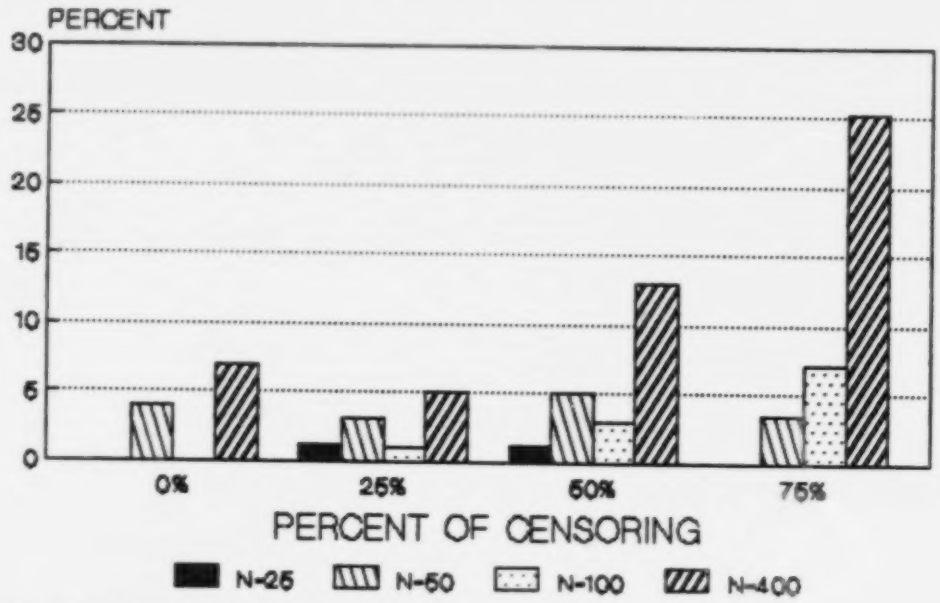
FIGURE 5a
PROPORTION OF CONGENERIC MODEL REJECTION
HIGH RELIABILITY MODEL Lambda = .90



Controlling for Nonconvergence
NTGLS ESTIMATION PROCEDURE

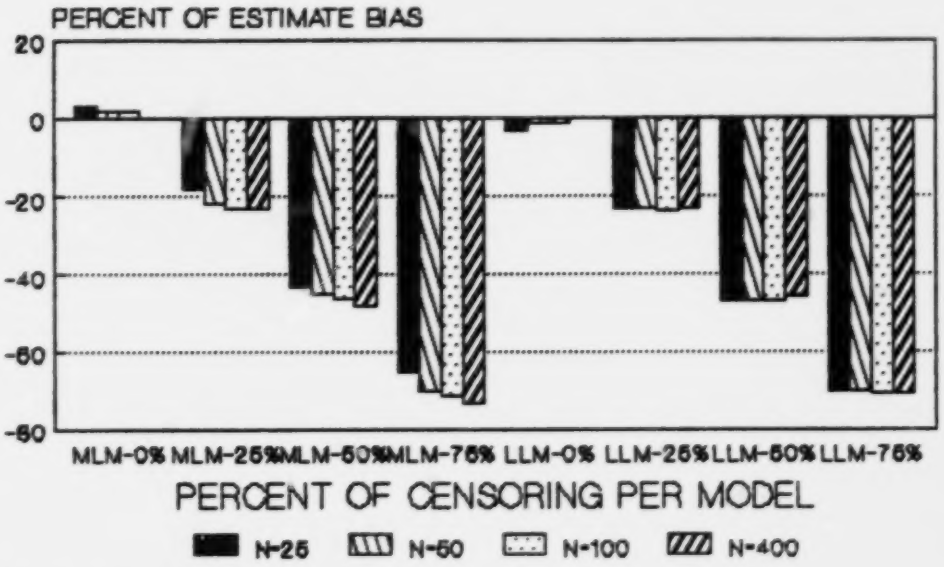
FIGURE 5b

PROPORTION OF CONGENERIC MODEL REJECTION
MODERATE RELIABILITY MODEL Lambda = .80



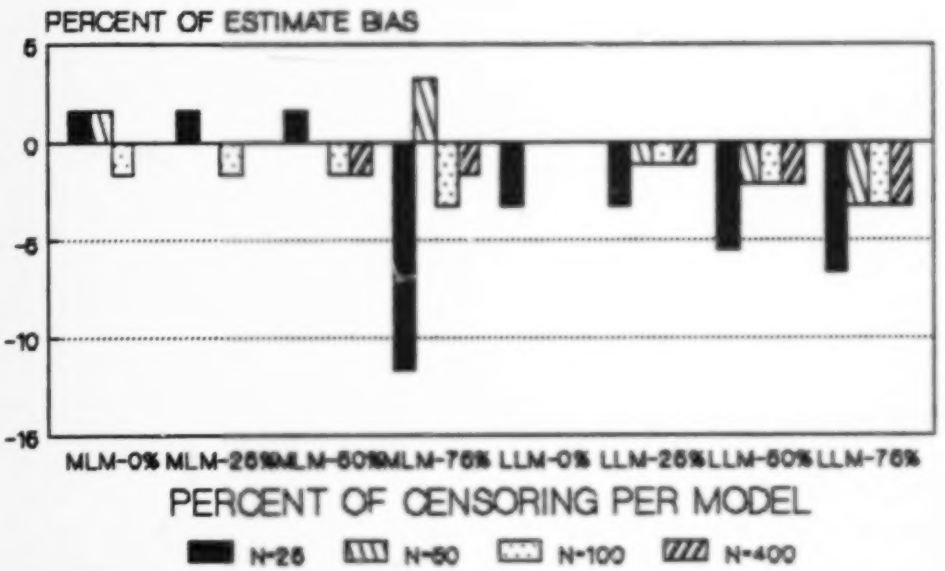
Controlling for Nonconvergence
NTGLS ESTIMATION PROCEDURE

FIGURE 6a
PERCENT OF BIAS IN LAMBDA (4,1) ESTIMATE
HIGH AND MODERATE RELIABILITY MODELS



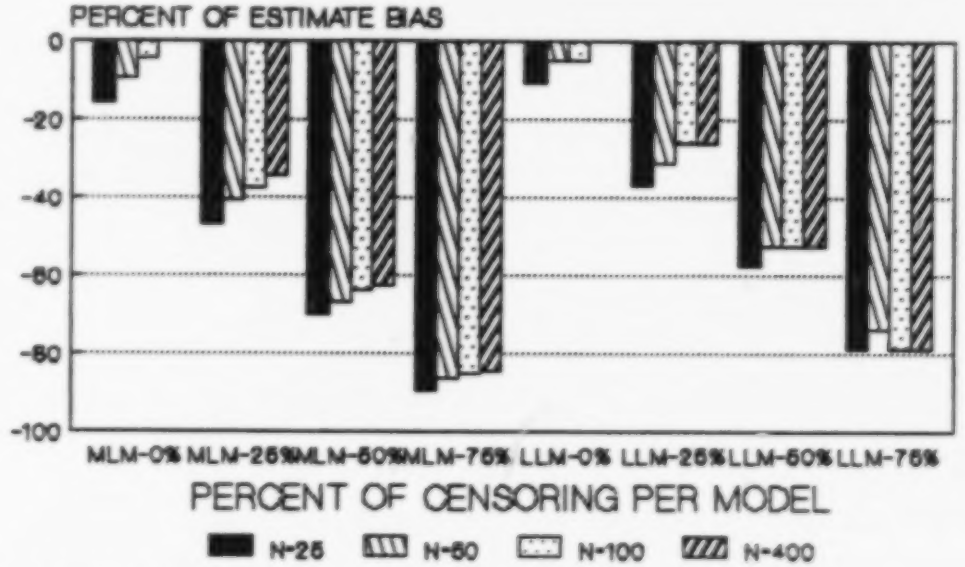
NTGLS ESTIMATION
MLM - Moderate Loadings Model .6
LLM - Large Loadings Model .9

FIGURE 6b
PERCENT OF BIAS IN LAMBDA (1,1) ESTIMATE
HIGH AND MODERATE RELIABILITY MODELS



NTGLS ESTIMATION
MLM - Moderate Loadings Model
LLM - Large Loadings Model

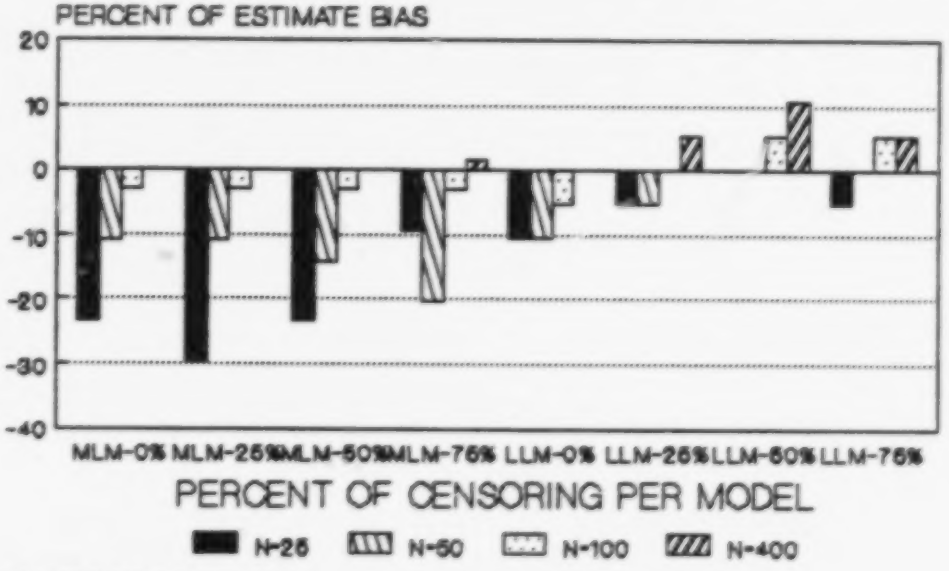
FIGURE 7a
PERCENT OF BIAS IN THETA-EPSILON (4,1)
HIGH AND MODERATE RELIABILITY MODELS



NTGLS ESTIMATION
MLM - Moderate Loadings Model
LLM - Large Loadings Model

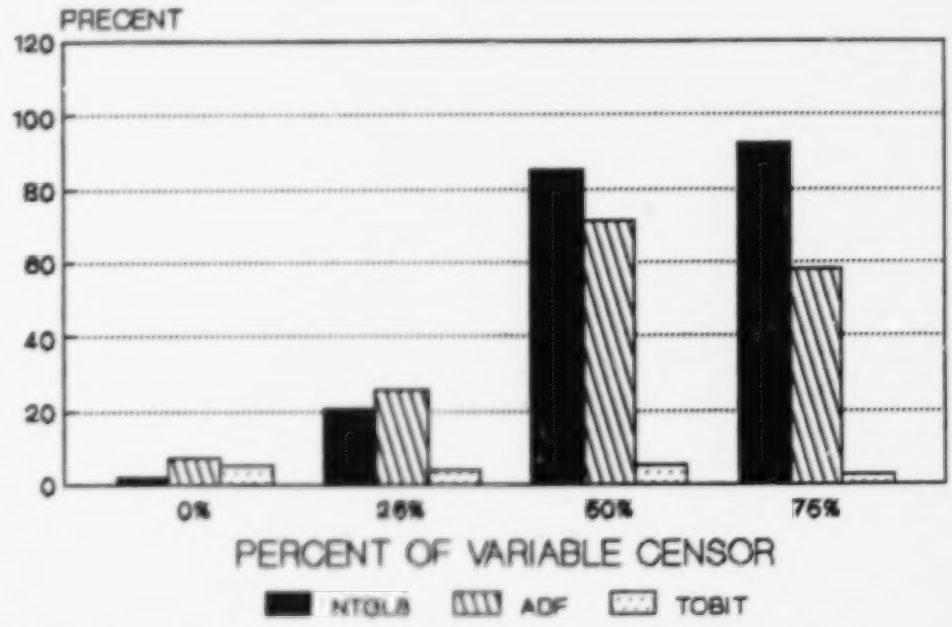
FIGURE 7b

PERCENT OF BIAS IN THETA-EPSILON (1,1)
HIGH AND MODERATE RELIABILITY MODELS



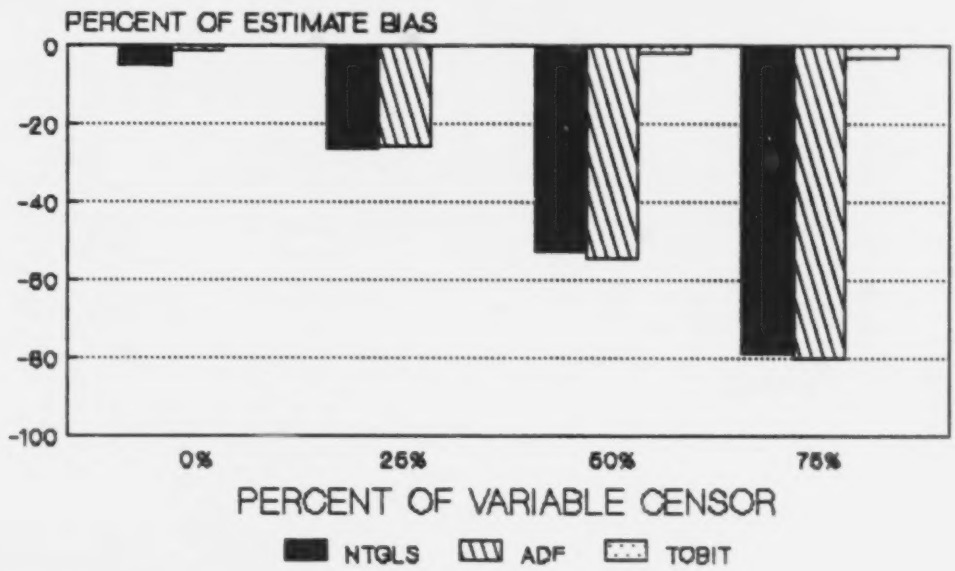
NTGLS ESTIMATION
MLM - Moderate Loadings Model
LLM - Large Loadings Model

FIGURE 8
PROPORTION OF MODEL REJECTIONS CONTROLLING FOR NONCONVERGENCE



Sample Size - 100

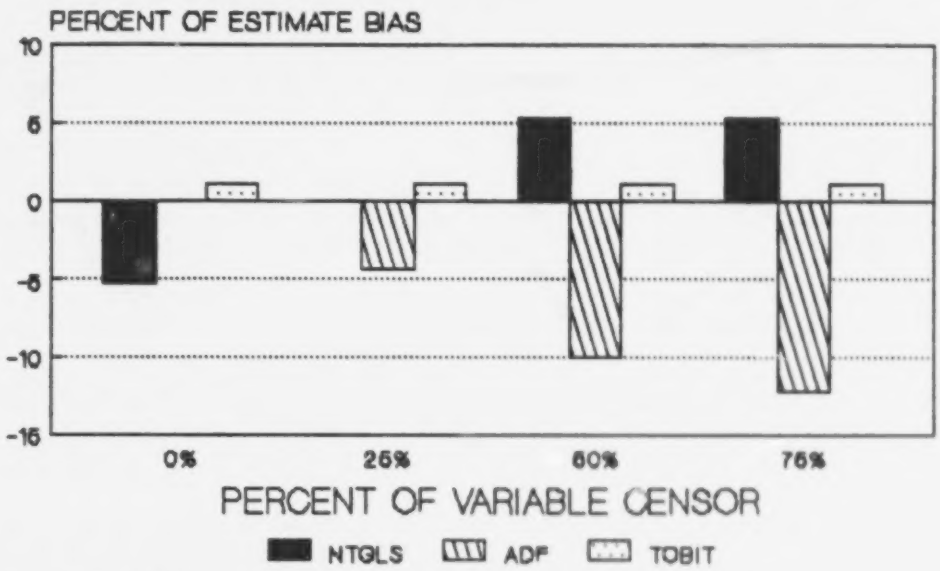
FIGURE 9a
PERCENT OF BIAS FOR LAMBDA (4,1)
SAMPLE SIZE = 100 OVER 100 REPLICATIONS



This may also be interpreted as the percent of bias in the reliability estimate for this rater.

FIGURE 9b

PERCENT OF BIAS FOR LAMBDA (1,1)
SAMPLE SIZE = 100 OVER 100 REPLICATIONS



This may also be interpreted as the percent of bias in the reliability estimate for this rater.

FIGURE 10a
PERCENT OF BIAS IN THETA-EPSILON (4,1)
ACROSS THE THREE ESTIMATION PROCEDURES

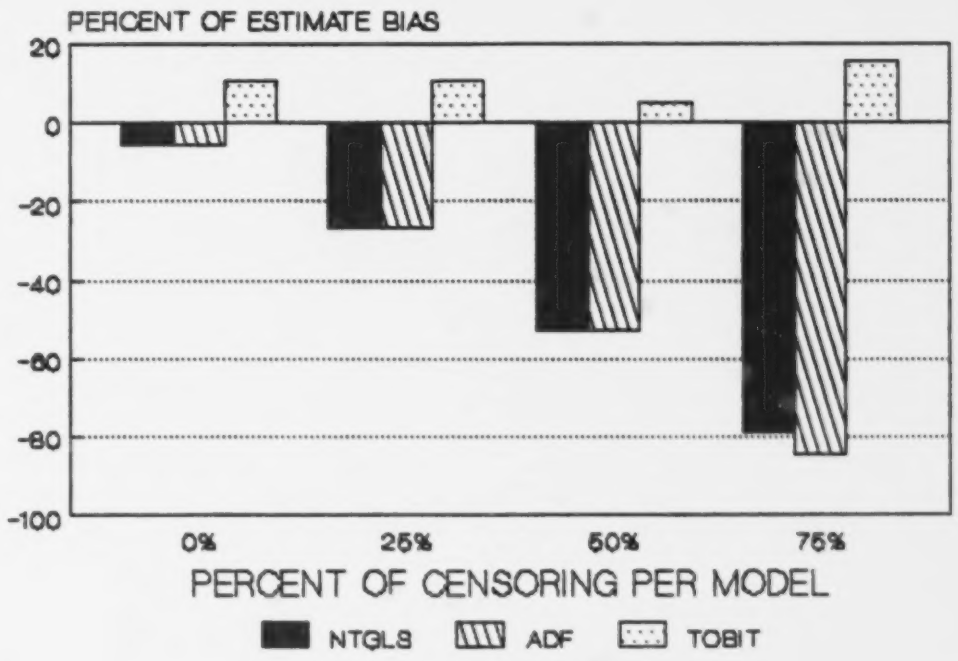


FIGURE 10b
PERCENT OF BIAS IN THETA-EPSILON (1,1)
ACROSS THE THREE ESTIMATION PROCEDURES